**LAPORAN PROYEK DATA MINING**

***Binary Classification using Random Forest***

****

**Disusun Oleh:**

12S18018 Yohana Polin Simatupang

12S18019 Maria Puspita Sari Nababan

12S18064 Letare Aiglien Saragih

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2021**

**DAFTAR ISI**

[**BAB 1 BUSINESS UNDERSTANDING** 5](#_Toc89760765)

[1.1 Determine Business Objective 5](#_Toc89760766)

[1.2 Determine Data Mining Goal 5](#_Toc89760767)

[1.3 Produce Project Plan 6](#_Toc89760768)

[**BAB 2 DATA UNDERSTANDING** 9](#_Toc89760769)

[2.1 Collect Initial Data 9](#_Toc89760770)

[2.2 Analysis Data 9](#_Toc89760771)

[2.3 Verify Data Quality 13](#_Toc89760772)

[2.4 Data Correlation 17](#_Toc89760773)

[**BAB 3 DATA PREPARATION** 21](#_Toc89760774)

[3.1 Sorting Data 21](#_Toc89760775)

[3.2 Cleaning Data 22](#_Toc89760776)

[3.3 Construct Data 23](#_Toc89760777)

[3.4 Binning 25](#_Toc89760778)

[3.5 Standardization 26](#_Toc89760779)

[**BAB 4 MODELLING** 28](#_Toc89760780)

[4.1 Build Test Scenario 28](#_Toc89760781)

[4.2 Model Building 30](#_Toc89760782)

[**BAB 5 MODEL EVALUATION** 33](#_Toc89760783)

[5.1 Evaluation of Modeling Result 33](#_Toc89760784)

[5.2 Modeling Process Review 34](#_Toc89760785)

[5.3 Determine Next Step 35](#_Toc89760786)

[**BAB 6 DEPLOYMENT** 36](#_Toc89760787)

[6.1 Model Deployment 36](#_Toc89760788)

[6.2 Final Report 36](#_Toc89760789)

[**LAMPIRAN** 38](#_Toc89760790)

**DAFTAR GAMBAR**

[Figure 1 Tahapan Knowledge Discovery in Databases (KDD) 6](#_Toc90930179)

[Figure 2 Memuat informasi ukuran data 9](#_Toc90930180)

[Figure 3 Memuat informasi bentuk data 9](#_Toc90930181)

[Figure 4 Proporsi kelas Label menggunakan histogram 10](#_Toc90930182)

[Figure 5 Informasi mengenai 53 fitur pada dataset 10](#_Toc90930183)

[Figure 6 Proporsi kelas Label menggunakan Pie Chart 11](#_Toc90930184)

[Figure 7 Potongan code untuk melihat tipe atribut 11](#_Toc90930185)

[Figure 8 Struktur dataset 13](#_Toc90930186)

[Figure 9 Potongan kode untuk melakukan imputasi data null dengan nilai mean 15](#_Toc90930187)

[Figure 10 Tampilan histogram untuk setiap fitur pada dataset 16](#_Toc90930188)

[Figure 11 Proporsi kelas severity level 17](#_Toc90930189)

[Figure 12 Korelasi antara fitur los dan umur 18](#_Toc90930190)

[Figure 13 Korelasi antara fitur los dan jnspelsep 18](#_Toc90930191)

[Figure 14 Korelasi antara fitur los dan sevetitylevel 19](#_Toc90930192)

[Figure 15 Korelasi setiap fitur pada dataset dengan heatmap 19](#_Toc90930193)

[Figure 16 Potongan code untuk menghapus fitur tertentu 21](#_Toc90930194)

[Figure 17 Potongan kode untuk melihat informasi atribut 21](#_Toc90930195)

[Figure 18 Informasi mengenai 53 fitur pada dataset 22](#_Toc90930196)

[Figure 19 Potongan kode untuk melihat missing value 22](#_Toc90930197)

[Figure 20 Transformasi atribut kategorik menjadi numerik 24](#_Toc90930198)

[Figure 21 Pengecekan atribut fitur setelah transforamasi 25](#_Toc90930199)

[Figure 22 Binning untuk fitur Umur 25](#_Toc90930200)

[Figure 23 Binning untuk fitur los 26](#_Toc90930201)

[Figure 24 Pembagian dan penyimpanand data dalam variabel X dan y 26](#_Toc90930202)

[Figure 25 Standarisasi fitur 27](#_Toc90930203)

[Figure 26 Perbandingan Hasil Akurasi pada Modelling Test 29](#_Toc90930204)

[Figure 27 Implementasi untuk membagi data menjadi data latih dan data uji 31](#_Toc90930205)

[Figure 28 Parameter setting 31](#_Toc90930206)

[Figure 29 Pemodelan dengan RFC 31](#_Toc90930207)

[Figure 30 Hasil akurasi data latih dan data uji 32](#_Toc90930208)

[Figure 31 Visualisasi hasil evaluasi dengan heatmap 34](#_Toc90930209)

[Figure 32 Rangkuman Tahapan Pengerjaan Proyek 37](#_Toc90930210)

**DAFTAR TABEL**

[Table 1 Tahap perencanaan yang dilakukan untuk mencapai tujuan data mining dan mencapai tujuan bisnis pada penelitian ‘Binary Classification using Random Forest’ 7](#_Toc90930291)

[Table 2 Informasi mengenai atribut, tipe atribut dan keterangan atribut 12](#_Toc90930292)

# **BAB 1 BUSINESS UNDERSTANDING**

Pada pengerjaan proyek ini akan dilakukan sesuai dengan tahapan pada metodologi CRISP DM yang akan dimulai dengan tahapan *business understanding* yaitu memahami permasalahan bisnis untuk proses *data mining* yang akan dilakukan. Adapun yang termasuk bagian dari tahapan ini adalah menentukan tujuan bisnis, menentukan sasaran yang ingin dicapai dengan data mining, dan menghasilkan perencanaan proyek yang akan dilakukan.

## **1.1 Determine Business Objective**

Rumah sakit merupakan salah satu instansi yang bergerak sebagai pelayanan kesehatan bagi masyarakat. Dalam melaksanakan proses bisnisnya, peran BPJS cukup besar dalam mempengaruhi kualitas pelayanan bagi masyarakat. Namun dengan semakin banyak penggunaan BPJS Kesehatan, tidak jarang terjadi beberapa kecurangan (*fraud*) yang ditujukan untuk menguntungkan pihak tertentu. Pelaku yang terlibat bisa jadi adalah peserta BPJS Kesehatan, *fasilitator* kesehatan atau pembeli layanan kesehatan, penyedia obat dan alat kesehatan, dan pemangku kepentingan lainnya. Penanganan terkait masalah tersebut menjadi *concern* yang perlu untuk diatasi yang bertujuan untuk dapat mencegah dan mendeteksi berbagai indikasi potensi kecurangan sedini dan sesedikit mungkin. Sehingga dengan demikian biaya pelayanan kesehatan dapat dimanfaatkan semaksimal mungkin dalam memenuhi kepentingan dan pelayanan yang maksimal bagi masyarakat, serta untuk tetap menjaga *sustainability* BPJS Kesehatan,

## **1.2 Determine Data Mining Goal**

Tujuan bisnis pada penelitian ini adalah untuk melakukan prediksi potensi terjadinya penyimpangan (*fraud*) pada klaim pelayanan Rumah Sakit. Melihat jumlah data yang besar dan studi kasus yang akan diteliti untuk itu, dilakukan penerapan *data mining* untuk menemukan pola menarik dari data. *Data mining* dikelompokkan menjadi *description, estimation, prediction, classification, clustering*, dan *association* [ref: buku pang-ning tan]*.* Pada penelitian ini, penggunaan data mining bertujuan sebagai dasar dalam pengembangan sebuah model klasifikasi biner untuk menemukan fraud. Ketika melakukan proses data mining, harus dilakukan beberapa tahapan antara lain, pembersihan data, integrasi data, pemilihan data, transformasi data, penemuan pola, evaluasi pola dan presentasi pengetahuan.

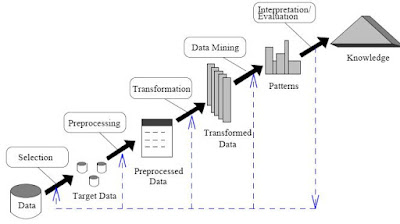


Figure 1 Tahapan Knowledge Discovery in Databases (KDD)

Dalam menemukan faktor apa saja yang menyebabkan terjadinya penyimpangan (*fraud*) pada layanan BPJS perlu digunakan data mining task dengan teknik asosiasi. *Association rule mining* adalah metode pembelajaran mesin berbasis aturan untuk menemukan hubungan yang menarik antara variabel dalam data yang berjumlah besar.

Algoritma yang akan untuk penelitian ini adalah Algoritma *Random Forest Classifier* (RFC). RFC merupakan metode klasifikasi yang *supervised* menggabungkan ratusan atau ribuan pohon keputusan, melatih masing-masing pohon pada serangkaian pengamatan yang sedikit berbeda, memisahkan simpul di setiap pohon dengan mempertimbangkan sejumlah fitur yang terbatas. Prediksi akhir dari random forest dibuat dengan merata-ratakan prediksi dari masing-masing pohon. Kelebihan dari algoritma ini adalah: menghasilkan eror yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data training dalam jumlah sangat besar secara efisien, efektif untuk mengestimasi hilangnya data, memperkiraan variabel apa yang penting dalam klasifikasi dan menyediakan metode eksperimental untuk mendeteksi interaksi variabel.

## **1.3 Produce Project Plan**

Tahap perencanaan yang dilakukan untuk mencapai tujuan data mining dan mencapai tujuan bisnis pada penelitian ‘*Binary Classification using Random Forest*’ ini adalah sebagai berikut:

Table 1 Tahap perencanaan yang dilakukan untuk mencapai tujuan data mining dan mencapai tujuan bisnis pada penelitian ‘Binary Classification using Random Forest’

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Aktivitas** | **Sub Aktivitas** | **Durasi** | **Sumber daya yang dibutuhkan** | **Ketergantungan** |
| Pemilihan Kasus dan Algoritma | Pemilihan Kasus | 1 | Semua analisis | - |
| Penentuan Algoritma | 6 |
| Business Understanding | Menentukan Objektif Bisnis | 1 | Semua analisis | Pemilihan kasus dan algoritma |
| Menentukan Tujuan Bisnis | 1 |
| Membuat Rencana Proyek | 1 |
| Data Understanding | Mengumpulkan Data | 1 | Semua analisis | Data dan teknologi |
| Menelaah Data | 1 |
| Memvalidasi Data | 1 |
| Data Preparation | Memilah Data | 1 | *Data mining consultant,* beberapa *database analyst time* | Data dan teknologi |
| mengkonstruksi Data | 4 |
| Menentukan Label Data | 1 |
| Membersihkan Data | 4 |
| Modeling | Membangun Skenario Pengujian | 3 | *Data mining consultant,* beberapa *database analyst time* | Algoritma |
| Membangun Model | 7 |
| Model Evaluation | Mengevaluasi Hasil Pemodelan | 5 | Semua analisis | Model yang telah dibuat |
| Melakukan Review Proses Pemodelan | 4 |
| Deployment | melakukan Deployment Model | 2 | *Data mining consultant,* beberapa *database analyst time* | Penerapan model berdasarkan data dan algoritma yang dipilih |
| Membuat laporan akhir Proyek | 4 |

Dalam pelaksanaan proyek dalam penelitian ini, diperlukan tools data mining yang mendukung metode untuk berbagai tahapan proses. Tools dan teknik yang digunakan dapat mempengaruhi keseluruhan proyek. Tools yang digunakan dalam mengerjakan proyek ini adalah python. Python adalah bahasa pemrograman berorientasi objek yang digunakan dalam pengembangan perangkat lunak maupun dalam analisis dan data science. Python memiliki berbagai library yang menyediakan fungsi untuk melakukan analisis data, memproses data, memvisualisasikan data, dll.

Berikut merupakan

# **BAB 2 DATA UNDERSTANDING**

Tahap kedua pada metodologi CRISP-DM setelah *business understanding* dalam melakukan metodologi *data science* adalah *data understanding*. Pada bab ini akan dijelaskan mengenai pengumpulan *initial data*, analisis untuk dapat memahami data yang akan digunakan dalam penelitian serta verifikasi pada kualitas data.

## **2.1 Collect Initial Data**

Langkah *data understanding* diawali dengan pengumpulan data yang akan digunakan pada proses *data science.* Data yang akan digunakan dalam kasus *binary classification* menggunakan *Random Forest Classification* (RFC) adalah data BPJS Kesehatan yang berasal dari dataset yang digunakan dalam kompetisi Hackathon. Data yang digunakan dalam penelitian ini merupakan data dengan format csv yang sudah terstruktur. Memuat informasi BPJS Kesehatan yang merupakan data publik mengenai aturan penamaan dan kesehatan secara umum. Data yang digunakan berukuran 10611501

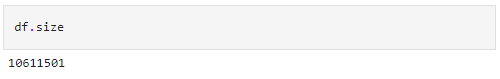


Figure 2 Memuat informasi ukuran data

## **2.2 Analysis Data**

Dataset train yang digunakan untuk memprediksi penyimpangan (fraud) pada layanan BPJS terdiri dari 200217 observasi dan 53 variabel dan memiliki proporsi kelas label pada data seimbang.

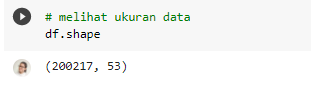


Figure 3 Memuat informasi bentuk data

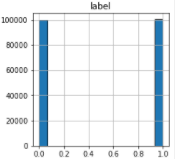


Figure 4 Proporsi kelas Label menggunakan histogram

Adapun ke 53 fitur/ variabel yang dimaksud adalah sebagai berikut:

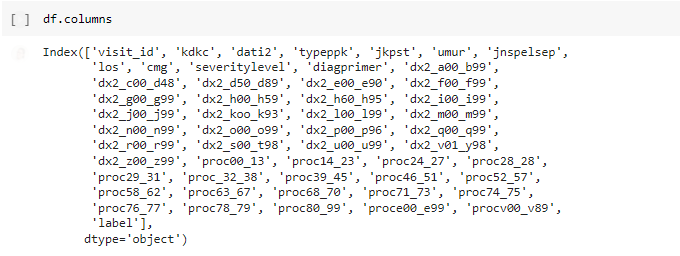


Figure 5 Informasi mengenai 53 fitur pada dataset

Pada tahap ini akan dilakukan pendefinisian label data yang akan digunakan. Tahap ini bertujuan untuk memastikan bahwa data yang digunakan merupakan *balance dataset*.

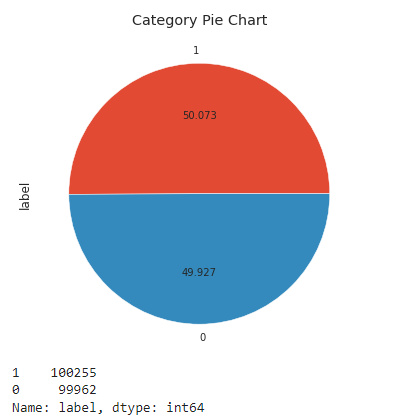


Figure 6 Proporsi kelas Label menggunakan Pie Chart

Berdasarkan pie chart diatas, dapat dilihat bahwa jumlah data yang menunjukkan *fraud* dan tidak *fraud* pada feature Label adalah seimbang.

Kemudian perlu dilakukan Exploratory Data Analysis (EDA). EDA digunakan untuk memahami data, mendapatkan konteks data, memahami variabel dan hubungan di antara variabel, dan merumuskan hipotesis yang berguna dalam membangun model prediksi. Langkah awal yg dilakukan untuk memahami data adalah dengan menganalisis tipe dari setiap fitur/ variabel yang akan digunakan menggunakan fungsi info().

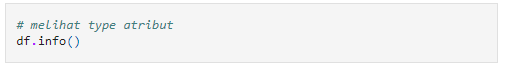


Figure 7 Potongan code untuk melihat tipe atribut

Berdasarkan fungsi tersebut diperoleh informasi mengenai tipe atribut atau fitur pada dataset sebagai berikut:

Table 2 Informasi mengenai atribut, tipe atribut dan keterangan atribut

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Variabel** | **Tipe Variabel** | **Deskripsi** |
| 1 | visit\_id | int64 | id kunjungan |
| 2 | kdkc | int64 | kode wilayah kantor cabang BPJS Kesehatan |
| 3 | dati2 | int64 | kode kabupaten/kota |
| 4 | typeppk | object | kode tipe Rumah Sakit |
| 5 | jkpst | object | jenis kelamin peserta JKN-KIS |
| 6 | umur | int64 | umur peserta saat mendapatkan pelayanan rumah sakit |
| 7 | jnspelsep | int64 | tingkat pelayanan; 1:rawat inap; 2: rawat jalan |
| 8 | los | int64 | lama peserta dirawat di rumah sakit |
| 9 | cmg | object | klasifikasi CMG (Case Mix Group) |
| 10 | severitylevel | int64 | tingkat urgensi |
| 11 | diagprimer | object | diagnosa primer |
| 12 | dx2\_...\_... | int64 | diagnosa sekunder |
| 13 | proc..\_... | int64 | kode kelompok procedure |
| 14 | label | int64 | flag fraud; 1:fraud; 0:tidak fraud |

Dari 53 fitur yang tersedia, terdapat 2 kategori yang diperoleh, yaitu: 4 fitur dengan data kategorik dan 49 fitur dengan data numerik. Untuk mendapatkan hasil analisa dataset yang lebih baik, maka perlu dilakukan pengidentifikasian kembali subset data yang relevan untuk kemudian digunakan pada tahapan selanjutnya yang sesuai dengan tujuan data mining pada penelitian ini.

## **2.3 Verify Data Quality**

Tahapan selanjutnya adalah melakukan verifikasi terhadap kualitas data yang digunakan. Untuk mendapatkan data yang berkualitas baik, perlu dilakukan pembersihan data (*dat cleaning*). Sebelum pembersihan data dilakukan, terlebih dahulu dilakukan pengecekan struktur data.

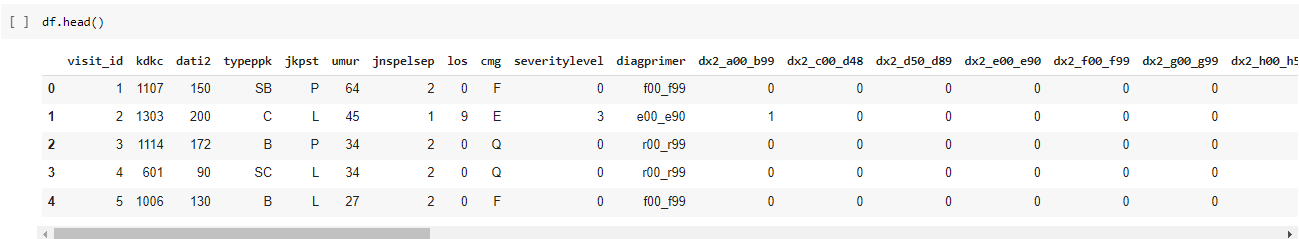
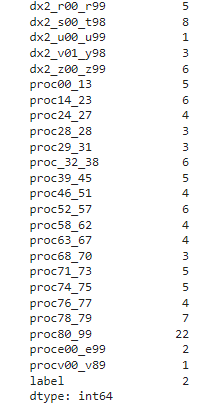
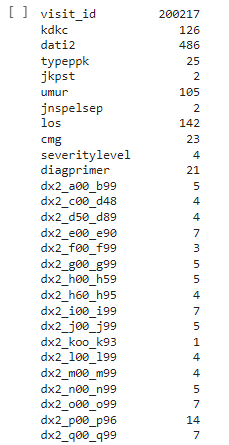
**

Figure 8 Struktur dataset

untuk mengetahui nilai unik di sepanjang sumbu (*axis*) kolom kita akan menggunakan fungsi nunique() yang akan memprint total nilai unik di setiap baris. Hal ini bertujuan untuk melihat kualitas nilai setiap fitur berdasarkan jumlah nilai pada tiap atribut.



Berdasarkan pengecekan diatas, dapat dilihat terdapat 3 atribut yang memiliki jumlah nilai = 1, yaitu: dx2\_koo\_k93, dx2\_u00\_u99 dan procv00\_v89.

Selanjutnya, dilakukan pengecekan kebersihan data dari kasus seperti *noisy*, *missing value*, dan masalah lainnya. *Data cleaning* pada proses data mining dapat mengurangi jumlah dan kompleksitas data. Salah satu aspek yang menyebabkan kualitas data menjadi kurang baik adalah terjadinya *missing value* atau terdapat data yang hilang pada dataset yang digunakan. Untuk mengantisipasi hal tersebut terlebih dahulu dilakukan pemeriksaan apakah terdapat data yang hilang (*missing*) atau bernilai kosong. Pemeriksaan dilakukan menggunakan fungsi pada python yaitu *df.isna()*

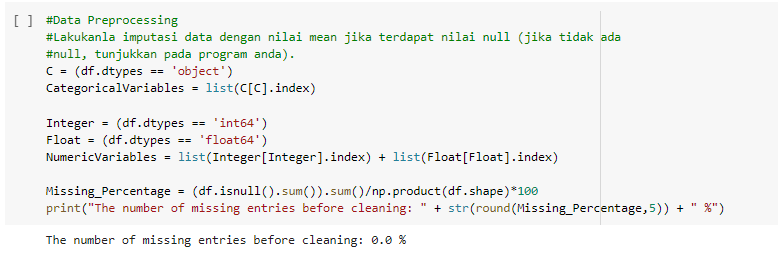
****

Figure 9 Potongan kode untuk melakukan imputasi data null dengan nilai mean

Adapun hasil yang didapatkan dari pemeriksaan tersebut adalah bahwa pada dataset yang digunakan tidak terdapat *missing value*.

Proses verifikasi kualitas data dilanjutkan dengan menggunakan visualisasi data dengan memanfaatkan fungsi hist untuk menampilkan histogram untuk semua atribut. Histogram dalam tampilan bentuk grafis akan menunjukkan distribusi data secara visual atau seberapa sering suatu nilai yang berbeda itu terjadi dalam suatu kumpulan data (*dataframe*). Histogram menunjukkan distribusi data dengan memplot frekuensi kejadian dalam suatu rentang.

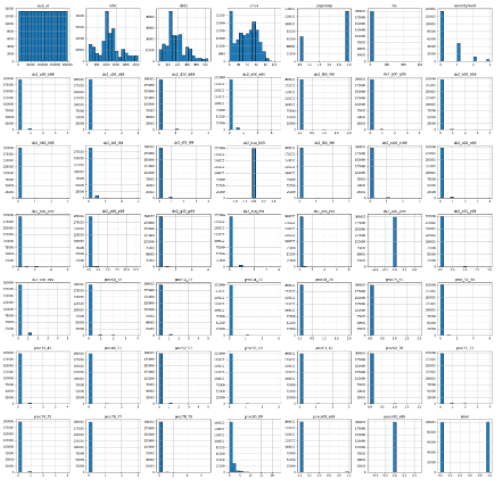


Figure 10 Tampilan histogram untuk setiap fitur pada dataset

Berdasarkan histogram yang dihasilkan, dapat dilihat bahwa kolom yang memiliki variasi data pada frekuensi tertentu adalah kdkc, dati2, dan umur. Namun persebaran atau distribusi data tidak tersebar secara konsisten. Kolom **kdkc** yang menunjukkan kode kantor cabang BPJS yang paling banyak adalah pada code di rentang 1000. Sementara pada kolom **dati2** yang menunjukkan kode kabupaten, paling tinggi berada pada rentang 100-200. Dan untuk kolom **umur**, nilai yang paling tinggi berada pada rentang umur 0.

## **2.4 Data Correlation**

Pada tahap ini, akan dilakukan pengecekan keterkaitan setiap fitur pada data yang digunakan untuk mengetahui bagaimana data akan dimanfaatkan untuk mengatasi masalah bisnis yang akan diselesaikan. Pada fitur severity level dimuat informasi mengenai tingkat urgensi rawat pasien yang dibagi menjadi 4 nilai yaitu 0-3. Urgensi kasus dalam INA-CBG terbagi menjadi:

1. “0” Untuk Rawat jalan
2. “I - Ringan” untuk rawat inap dengan tingkat keparahan 1 (tanpa komplikasi maupun komorbiditi)
3. “II - Sedang” Untuk rawat inap dengan tingkat keparahan 2 (dengan mild komplikasi dan komorbiditi)
4. “III - Berat” Untuk rawat inap dengan tingkat keparahan 3 (dengan major komplikasi dan komorbiditi

Berikut merupakan pie chart yang menampilkan perbandingan dari keempat nilai urgensi tersebut:

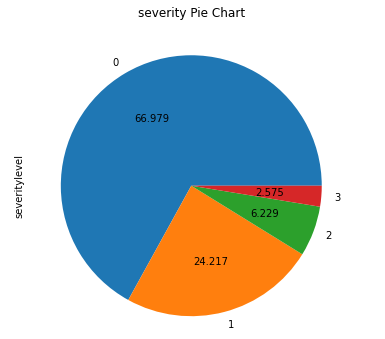


Figure 11 Proporsi kelas severity level

Berdasarkan output tersebut dapat diketahui bahwa nilai paling tinggi ditunjukkan oleh kelas 0 yaitu "rawat jalan". Sub-group tersebut merupakan resource intensity level yang menunjukkan tingkat keparahan kasus yang dipengaruhi adanya komorbiditas ataupun komplikasi dalam masa perawatan.

Kedua adalah melihat korelasi antara fitur los dan umur. Hal ini bertujuan untuk melihat umur pasien yang lebih banyak dirawat.

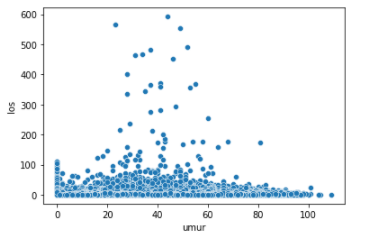


Figure 12 Korelasi antara fitur los dan umur

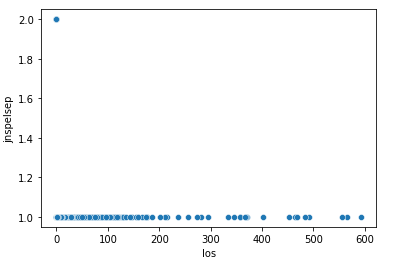


Figure 13 Korelasi antara fitur los dan jnspelsep

Pada atribut jnspelsep terdapat 2 jenis nilai yang digunakan, yaitu 1 dan 2. Nilai 1 diartikan sebagai pasien yang mendapat layanan rawat inap dan nilai 2 diartikan sebagai pasien yang mendapat layanan rawat jalan.

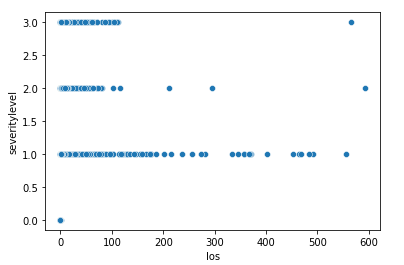


Figure 14 Korelasi antara fitur los dan sevetitylevel

Setelah melakukan proses binning terhadap length of stay (los) yang dibagi menjadi rawat jalan, short stay, medium stay, dan long stay. Dapat dilihat terdapat korelasi tidak valid, saat dirawat jalan maka harusnya rawat jalan hanya ada pada koordinat 2, namun dari hasil visualisasi terdapat tingkat pelayanan rawat jalan yang los nya lebih dari 0 hari, atau menginap.

Keterkaitan (korelasi) tersebut dapat dilihat dengan memvisualisasikan data menggunakan heatmap ataupun scatter plot.

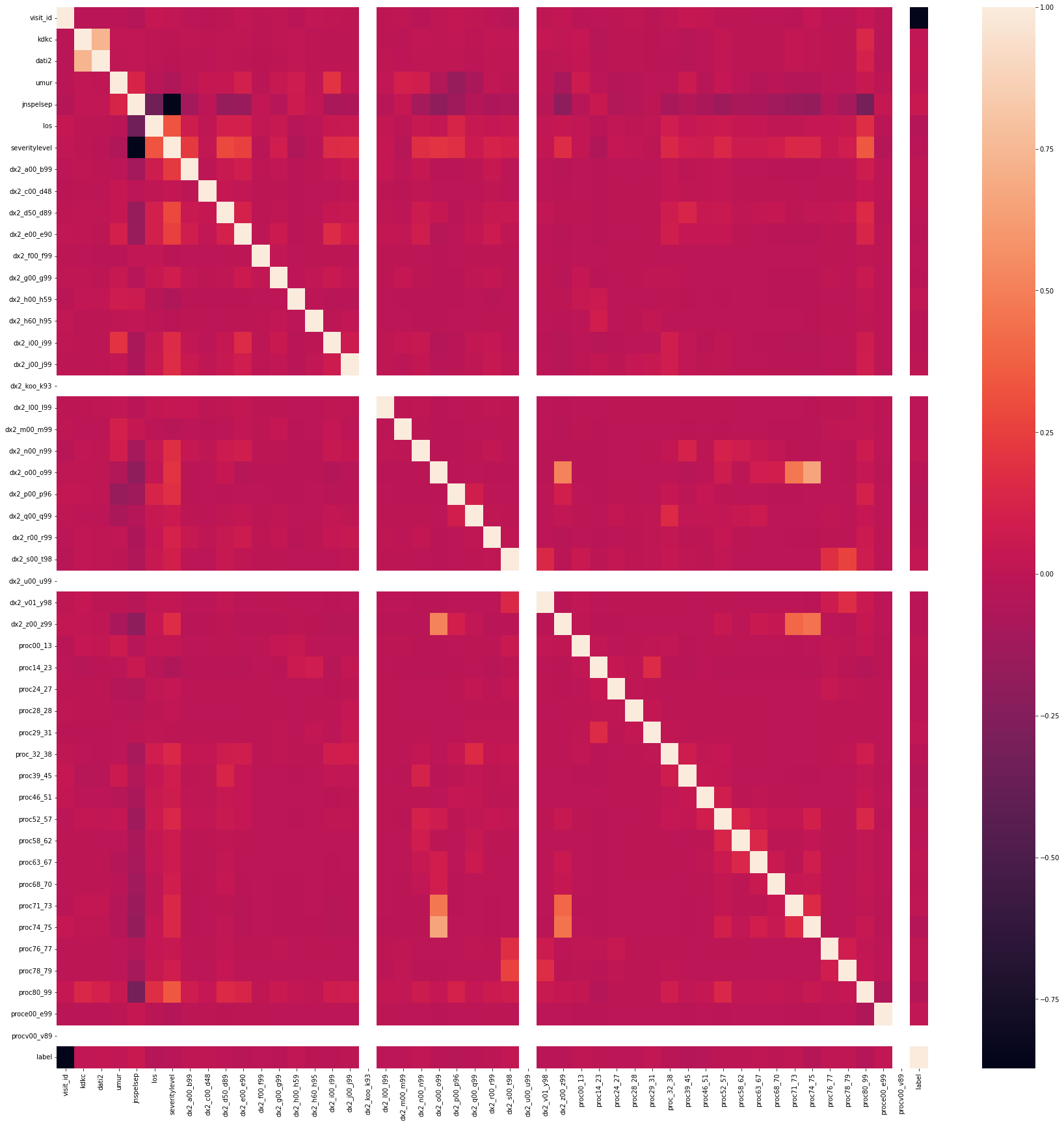


Figure 15 Korelasi setiap fitur pada dataset dengan heatmap

Pada gambar dapat dilihat bahwa 3 atribut dengan jumlah nilai =1 seperti yang disebutkan pada tahap verifikasi data, yaitu: dx2\_koo\_k93, dx2\_u00\_u99 dan procv00\_v89, tidak memiliki korelasi dengan fitur lain. Sementara 50 fitur lainnya berkorelasi satu sama lain dengan tingkat ketergantungan yang berbeda.

# 

# **BAB 3 DATA PREPARATION**

Tahap ketiga pada metodologi CRISP-DM setelah *data understanding* dalam melakukan metodologi *data science* adalah *data preparation*. Pada bab ini akan dijelaskan mengenai proses apa saja yang akan dilakukan untuk mempersiapkan data seperti *sorting, cleaning, construction, binning* dan *normalization.*

## **3.1 Sorting Data**

Data yang akan digunakan dalam proses *data mining* terlebih dahulu perlu dipersiapkan dengan baik. Fase *sorting* merupakan tahapan untuk melakukan pemilihan pada atribut yang akan digunakan. Atribut yang tidak digunakan akan *di drop*.



Figure 16 Potongan code untuk menghapus fitur tertentu

Atribut tersebut di *drop* dengan tujuan agar data yang digunakan lebih efisien dan efektif dalam pengolahan data termasuk dalam penggunaan memory. Berikut adalah tampilan setelah atribut yang tidak digunakan telah di *drop.*

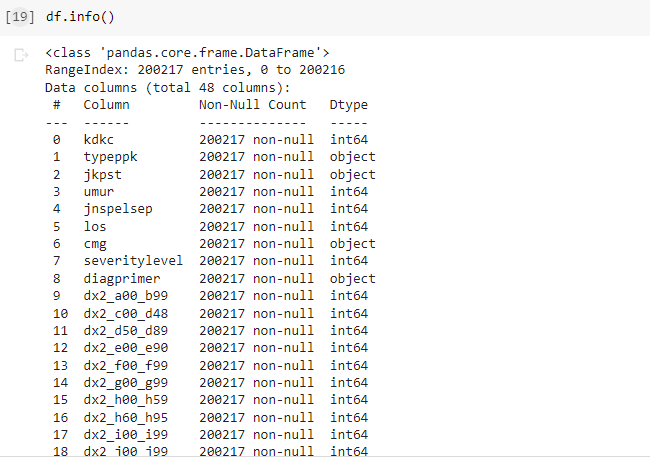
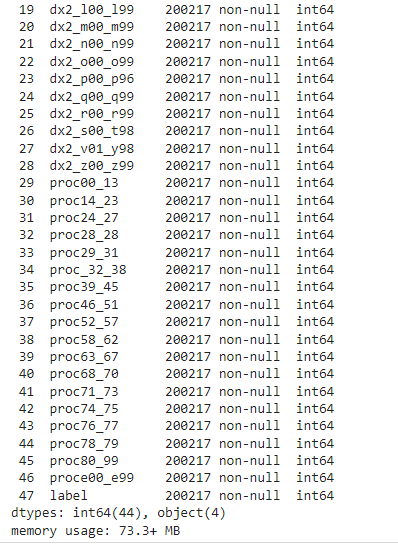
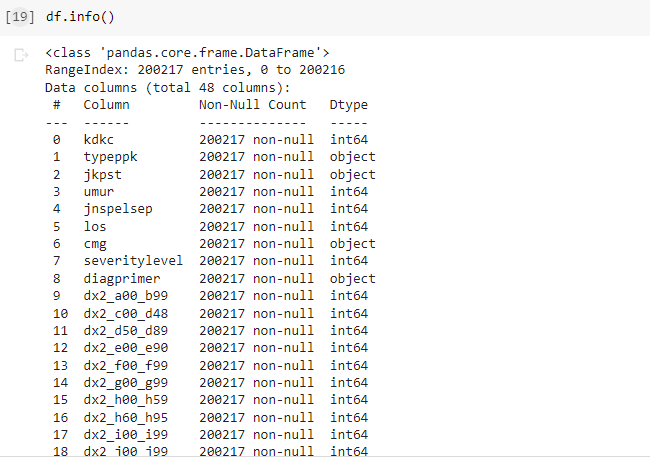


Figure 17 Potongan kode untuk melihat informasi atribut



Nilai penggunaan *memory* menjadi berkurang setelah dilakukan pemilihan atribut yang diperlukan yaitu sebagai berikut

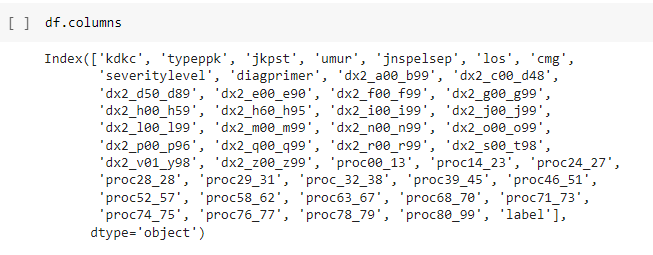


Figure 18 Informasi mengenai 53 fitur pada dataset

## **3.2 Cleaning Data**

Fase ini merupakan tahapan untuk melakukan pembersihan data. Pembersihan data yang dilakukan adalah menangani objek data yang kosong (*missing value*). Untuk itu, terlebih dahulu dilakukan pemeriksaan data untuk memeriksa apakah terdapat nilai yang hilang (*missing*)

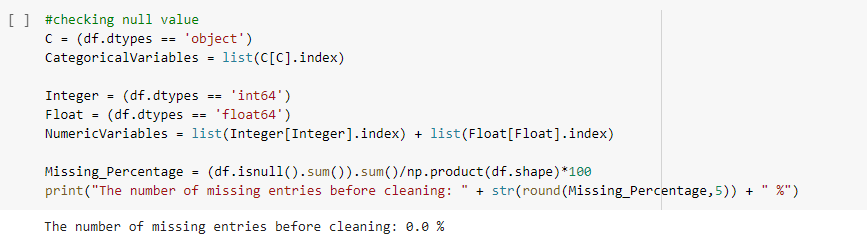
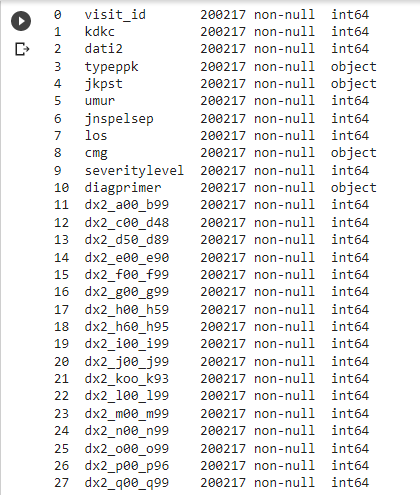


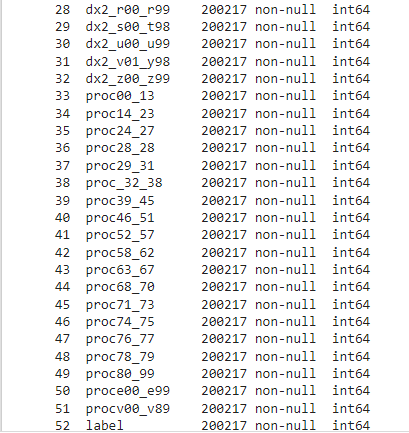
Figure 19 Potongan kode untuk melihat missing value

Python Pandas memungkinkan kita dapat menemukan *missing value* secara cepat dengan fungsi isna().Fungsi isna() akan mengembalikan nilai boolean dari dataset yang diperiksa. Hasil keluaran berupa **False** menunjukkan bahwa pada cell tersebut tidak terdapat nilai yang kosong (*missing*). Agregasi data dengan fungsi sum() ditujukan agar dapat memahami data dengan lebih baik. Agregasi sum()akan menjumlahkan semua cell yang kosong apabila terdapat nilai yang kosong pada atribut tertentu.

## **3.3 Construct Data**

Fase ini merupakan tahapan untuk melakukan konstruksi pada data. Adapun konstruksi yang dilakukan adalah transformasi atribut dengan tipe kategorik menjadi numerik. Hal ini bertujuan agar data kemudian dapat di normalisasi. Untuk tahap pada konstruksi data dilakukan pengecekan tipe data pada dataset menggunakan fungsi df.info(), dan output yang dihasilkan adalah sebagai berikut:





Dapat dilihat pada gambar di atas, terdapat 4 atribut yang bertipe data kategorikal (object64), untuk itu perlu dilakukan transformasi data. Untuk itu perlu dilakukan transformasi data tipe pada atribut dengan menjalankan potongan kode berikut:

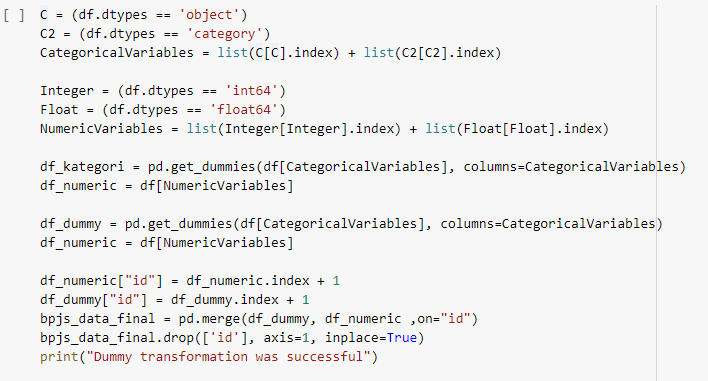


Figure 20 Transformasi atribut kategorik menjadi numerik

Setelah transformasi berhasil, dilakukan pengecekan kembali pada type atribut menggunakan fungsi df.info()

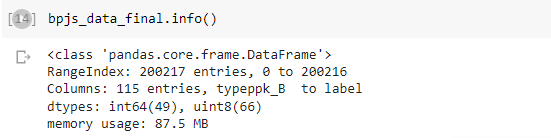


Figure 21 Pengecekan atribut fitur setelah transforamasi

## **3.4 Binning**

Tahapan ini merupakan proses transformasi data dengan menggunakan metode *binning.* Metode ini akan digunakan untuk mengelompokkan data numerik menjadi beberapa bin dengan tujuan memudahkan pemahaman pada persebaran data yang digunakan. Berdasarkan analisis yang didapatkan, diketahui bahwa fitur **umur** dan **LoS** merupakan nilai bertipe numerik dan memiliki persebaran data yang tidak merata. Oleh karena itu akan dilakukan proses *binning* pada kedua fitur tersebut.

Untuk fitur umur akan dibagi menjadi 5 kategori dengan bin yang ditentukan adalah sesuai dengan kategori usia berdasarkan WHO yaitu sebagai berikut.

Bin 1: umur <=1,

Bin 2: 2<= umur >=10,

Bin 3: 11<= umur >=19,

Bin 4: 20<= umur >=60,

Bin 5: umur > 60

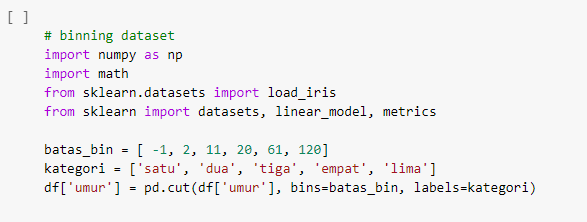


Figure 22 Binning untuk fitur Umur

Untuk fitur los yang memiliki hubungan terhadap jnpsplesep yang terkait pada tipe rawat inap atau rawat jalan selanjutnya akan dikelompokkan menjadi 4 kategori yaitu ‘rawat jalan’, ‘short stay’, ‘medium stay’, ‘long stay’. Penentuan bin adalah sebagai berikut.

los = 0

0 : rawat jalan,

1-5 : short stay,

6- 10 : medium stay,

> 10 : long stay



Figure 23 Binning untuk fitur los

## **3.5 Standardization**

Untuk tahap ini akan dilakukan standarisasi pada data yang telah diolah sebelumnya untuk mendapatkan hasil yang lebih baik. Sebelum standarisasi dijalankan, terlebih dahulu data dibagi dan disimpan dalam variabel X dan y seperti dibawah ini.

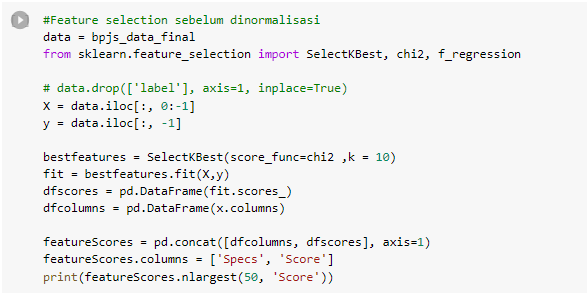


Figure 24 Pembagian dan penyimpanand data dalam variabel X dan y

Penerapan standarisasi berfokus pada mengubah data mentah menjadi informasi yang dapat digunakan sebelum dianalisis. Merupakan teknik yang menskalakan data sehingga memiliki mean = 0 dan standar deviasi =1.

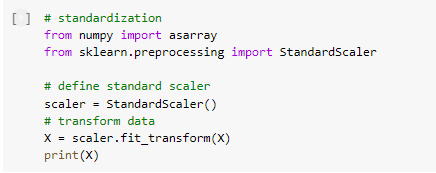
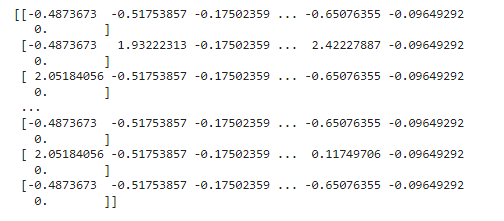


Figure 25 Standarisasi fitur

Dan output yang diperoleh adalah sebagai berikut:



# **BAB 4 MODELLING**

Tahap keempat pada metodologi CRISP-DM untuk melakukan binary classification dalam mendeteksi fraud adalah modeling. Pada bab ini akan dijelaskan mengenai pemilihan teknik modelling, dan menghasilkan *test scenario* serta teknik membangun model yang akan dibangun.

## **4.1 Build Test Scenario**

Pada proses melakukan *data mining*, pemilihan model akan dipengaruhi oleh tujuan dari pelaksanaannya. Sebelum melakukan pembangunan model, perlu dilakukan perancangan bagaimana model yang akan dibangun. Analisis melalui pengujian model yang akan dipilih yaitu sebagai berikut.

1. Model menggunakan seluruh features

Pada model ini, akan dibangun menggunakan seluruh features pada dataset. Sebelumya diketahui terdapat 53 features sebelum dilakukan *data preprocessing*. Pada model ini akan dilakukan prediksi menggunakan RandomForestClassification. Prediksi yang dilakukan menghasilkan akurasi untuk data train dan data test masing-masing sebesar 0.93 dan 0.67

1. Model menggunakan best features

Pada teknik pemodelan berikutnya, dilakukan pemilihan *best features* dengan memanfaatkan fungsi SelectKBest dengan total K sebesar 70. Pada model ini akan dilakukan prediksi menggunakan RandomForestClassification. Prediksi yang dilakukan menghasilkan akurasi untuk data train dan data test masing-masing sebesar 0.85 dan 0.67

1. Model menggunakan fitur ['kdkc', 'typeppk', 'jkpst', 'umur', 'jnspelsep', 'los', 'cmg', 'severitylevel', 'diagprimer', 'label']

Pada pemodelan ini akan dilakukan dengan memilih hanya feature tertentu untuk digunakan sebagai data test maupun data train. Untuk itu, features yang tidak digunakan akan di drop sesuai dengan kebutuhan. Adapun features yang akan digunakan adalah 'kdkc', 'typeppk', 'jkpst', 'umur', 'jnspelsep', 'los', 'cmg', 'severitylevel', 'diagprimer', 'label'. Pada model ini akan dilakukan prediksi menggunakan RandomForestClassification. Prediksi yang dilakukan menghasilkan akurasi untuk data train dan data test masing-masing sebesar 0.89 dan 0.67

1. Model dengan menggunakan tuning hyperparameter pada fitur yang dipilih

Pada pemodelan ini akan dilakukan teknik tuning hyper parameter yaitu kita dapat melakukan pengaturan pada algoritma dengan mengubah parameter untuk menemukan kinerja yang optimal. Jumlah *estimator* akan ditunjukkan dengan nilai *start* dan *stop* yang telah ditentukan, kemudian hyperparameter akan ditemukan dengan menggunakan fungsi random\_grid. Hasil hyperparameter tersebut kemudiana akan digunakan sebagai parameter prediksi dengan fungsi RandomizedSearchCV. Prediksi yang dilakukan menghasilkan akurasi untuk data train dan data test masing-masing sebesar 0.81 dan 0.68

Berikut merupakan tampilan visualisasi yang menampilkan hasil akurasi yang diperoleh dari masing-masing pemodelan yang dilakukan

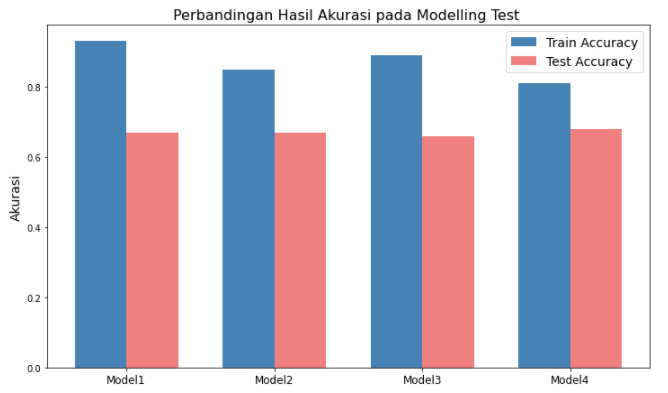


Figure 26 Perbandingan Hasil Akurasi pada Modelling Test

Berdasarkan viasualisasi diatas, Model1 merupakan pemodelan menggunakan semua fitur pada dataset, Model2 merupakan pemodelan menggunakan best *feature*, Model 3 merupakan pemodelan menggunakan fitur tertentu, dan Model4 merupakan pemodelan menggunakan *Tuning Hyperparameter*. Dapat dilihat hasil akurasi pemodelan tertinggi untuk data latih diperoleh menggunakan pemodelan jenis 1 yaitu menggunakan semua fitur, sementara hasil akurasi pemodelan tertinggi untuk data uji diperoleh dengan menggunakan pemodel jenis 4 yaitu *Tuning Hyperparameter*.

## **4.2 Model Building**

Berdasarkan pengujian untuk model yang telah ditemukan sebelumnya, maka pada proyek ini akan menggunakan model dengan tuning hyperparameter untuk *features* yang dipilih. Hal ini berdasarkan hasil yang diperoleh dari akurasi untuk *data train* dan *data test* yang menunjukkan *overfitting* yang lebih kecil dibandingkan dengan model lainnya. Dalam pembangunan model klasifikasi terdapat 3 informasi yang perlu didefinisikan untuk kemudian digunakan dalam pengambilan keputusan dalam *data mining*, antara lain:

1. Parameter *settings,* digunakan untuk penentuan parameter yang akan digunakan pada model
2. Membuat model menggunakan algoritma yang sudah ditentukan
3. Menampilkan hasil penilaian akurasi terhadap data latih dan data uji yang dimiliki

*Binary Classification* dengan algoritma RFC dibangun pada bahasa pemrograman *python* dengan memanfaatkan *library python* yaitu *scikit-learn*. *Scikit-learn* merupakan salah satu *library* yang disediakan *python* untuk membangun model machine learning seperti regresi, *clustering* dan *classification*. Pada tahap pemodelan ini, dataset yang digunakan merupakan dataset yang telah diproses sebelumnya seperti yang sudah dijelaskan pada bab 2 dan 3. Untuk pengimplementasian model RFC, tahap pertama yang dilakukan adalah membagi 2, yaitu: data latih dan data uji dengan persentase 70% untuk data latih dan 30% untuk data uji. Data latih akan digunakan untuk membangun model dan data uji akan digunakan untuk menguji model yang telah dibangun.

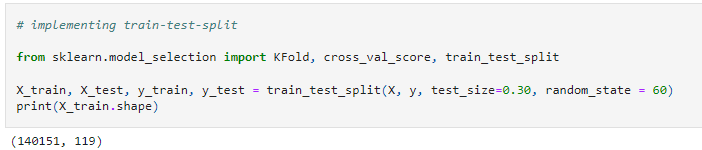


Figure 27 Implementasi untuk membagi data menjadi data latih dan data uji

Kemudian selanjutnya dilakukan pendefinisian 3 informasi yang dibutuhkan dalam pembuatan keputusan *data mining*, yaitu:

1. Parameter *settings,* digunakan untuk penentuan parameter yang akan digunakan pada model

Berdasarkan pengujian parameter yang telah dilakukan dengan parameter tuning, diperoleh kesimpulan bahwa parameter random\_state=5, n\_estimators=20 menghasilkan pemodelan dengan akurasi terbaik. Maka pada pemodelan RFC parameter random\_state = 5 dan n\_estimators = 20 akan digunakan.



Figure 28 Parameter setting

1. Membuat model menggunakan algoritma yang sudah ditentukan

Selanjutnya adalah pembangunan model berdasarkan algoritma yang dipilih yaitu RFC. Untuk pembangunan model sendiri menggunakan potongan kode berikut ini:

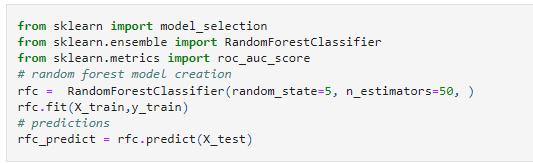


Figure 29 Pemodelan dengan RFC

1. Menampilkan hasil penilaian akurasi terhadap data latih dan data uji yang dimiliki

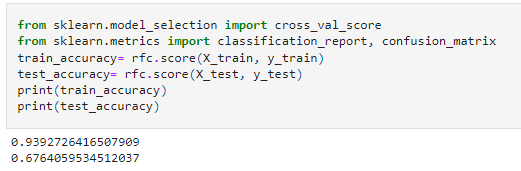


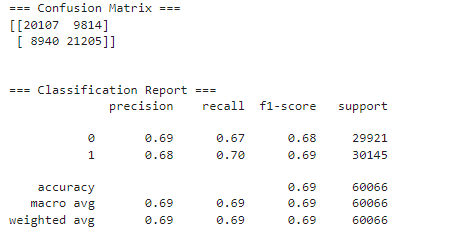
Figure 30 Hasil akurasi data latih dan data uji

# **BAB 5 MODEL EVALUATION**

Pada bab ini akan dijelaskan mengenai evaluasi terhadap model pendeteksi potensi kecurangan pada layanan BPJS yang dihasilkan menggunakan algoritma *Random Forest Classification.* Evaluasi adalah fase interpretasi terhadap hasil *data mining*. Evaluasi dilakukan secara mendalam dengan tujuan agar hasil pada tahap *modelling* sesuai dengan sasaran yang ingin dicapai.

## **5.1 Evaluation of Modeling Result**

Tahap ini dilakukan untuk mengetahui performa *binary classification* untuk mendeteksi *fraud* menggunakan *confusion matrix* dan *classification report* berdasarkan dataset yang digunakan yaitu data BPJS Kesehatan yang berasal dari dataset yang digunakan dalam kompetisi Hackathon. Sebelum pengerjaan proyek telah ditetapkan serangkaian ketentuan/ standar akurasi *precision, recall* dan *accuracy* pembangunan model. Dimana score *precision* > 0.54*, recall* > 0.65 dan *accuracy* > 0.56*.* Pada tahap pembangunan model, telah dilakukan penilaian akurasi terhadap data latih dan data uji. Dan pada tahap ini dilakukan evaluasi pemodelan dengan melihat *precision, recall* dan *accuracy* yang dilakukan adalah sebagai berikut:



Berdasarkan hasil yang diperoleh dari pembangunan model dengan menggunakan algoritma RFC telah menghasilkan model dengan akurasi cukup baik dengan score > 0.5 dan yang memenuhi standar dan ketentuan pembangunan proyek. Model yang dibangun telah cukup baik dalam menerapkan algoritma RFC untuk mendeteksi kecurangan pada layanan BPJS.

Selanjutnya evaluasi dilanjutkan dengan melakukan pemetaan kesesuaian output dari model menggunakan visualisasi heatmap, dan diperoleh hasil sebagai berikut:

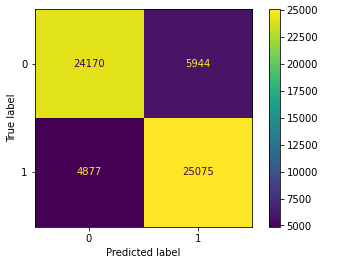


Figure 31 Visualisasi hasil evaluasi dengan heatmap

Karena penelitian ini merupakan *binary classification*, maka output akhir dari pemodelan ini adalah binary [0, 1], dimana 0 berarti terdapat tidak *fraud* dan 1 artinya terdapat *fraud*. Berdasarkan heatmap yang diperoleh dapat dilihat hubungan *predicted lable* dengan *true lable* dalam menghasilkan data *valid* dan tidak *valid*. Data valid yang diperoleh berupa: data yang diprediksi tidak *fraud* dan benar tidak *fraud* berjumlah 24170 dan data yang diprediksi *fraud* dan benar *fraud* berjumlah 25075. Sementara untuk data tidak *valid* yang diperoleh berupa: data yang diprediksi tidak *fraud* tetapi kebenarannya adalah *fraud* berjumlah 5944 dan data yang diprediksi *fraud* tetapi kebenarannya adalah tidak *fraud* berjumlah 4877.

## **5.2 Modeling Process Review**

Tahap ini memeriksa kembali tahapan dari awal untuk memastikan bahwa tidak ada faktor penting dalam proses tersebut yang terabaikan atau terlewati. Berdasarkan hasil peninjauan proses awal proyek data mining dengan metodologi CRISP-DM, maka dapat dipahami bahwa:

* Proses eksplorasi data akan membantu dalam memilih atribut yang berkaitan dengan mendeteksi terjadinya *fraud* pada layanan BPJS.
* *Data Preparation*, khususnya pada proses data *cleaning* dan *transform*, sehingga data yang diperoleh dapat menghasilkan model yang baik.
* Sangat penting untuk tetap fokus pada masalah bisnis yang dihadapi, karena setelah data siap dianalisis, maka akan dilakukan tahap pemodelan. *Business understanding* sangat penting dalam memutuskan bagaimana menerapkan hasil yang diperlukan dalam mendeteksi terjadinya *fraud* pada layanan BPJS.

## **5.3 Determine Next Step**

Tahapan ini menentukan langkah apa yang akan diambil selanjutnya. Berdasarkan hasil evaluasi terhadap model yang digunakan dengan algoritma RFC, dengan hasil akurasi pemodelan yang diperoleh dalam mendeteksi terjadinya *fraud* pada layanan BPJS, maka diputuskan pengerjaan proyek akan dilanjutkan ke tahap akhir yakni deployment.

## 

# **BAB 6 DEPLOYMENT**

Tahap keenam pada metodologi CRISP-DM untuk melakukan prediksi kinerja karyawan adalah deployment. Pada bab ini akan dijelaskan mengenai perencanaan dan *deployment* model yang sudah dihasilkan, serta laporan akhir untuk proses *data mining* yang sudah dilakukan.

## **6.1 Model Deployment**

Model yang sudah selesai dibangun selanjutnya dilanjutkan pada tahap *deployment.* Model *deployment* merupakan proses dimana model yang telah dibangun akan tersedia pada lingkungan produksi dimana model tersebut dapat melakukan prediksi pada sistem lain. Model *deployment* yang dilakukan pada proyek ini adalah berdasarkan pola secara dinamis yang akan di-*deploy* pada web browser, sehingga akan ditampilkan dalam bentuk website. *Deployment* model yang telah dibangun akan dilakukan pada aplikasi Heroku yaitu salah satu tools yang termasuk pada *Platform As A Service (PaaS)* untuk mengelola dan menjalankan aplikasi dari model yang dikembangkan. Aplikasi tersebut akan diterapkan pada Heroku dengan menggunakan Flask Python

## **6.2 Final Report**

Selama pengerjaan proyek ini, anggota tim terlibat dan berkontribusi dalam pengerjaan proyek dari awal hingga tahapan selesai dilakukan. Hal ini menjadi sarana pembelajaran bagi anggota tim menerapkan dalam dunia nyata pelaksanaan *data mining* sesuai dengan tahapan CRISP-DM. Tim juga dapat memahami dan mampu bereksplorasi pada data, tahap pemrosesan data, penerapan algoritma dalam membangun model, melakukan evaluasi untuk menilai performa model, hingga melakukan *deployment* untuk model yang telah dibangun. Tahapan yang dilakukan tim proyek setelah melakukan *deployment* adalah membuat dokumentasi yang dituangkan dalam laporan akhir. Laporan akhir mencakup penjelasan terkait dengan rangkaian proses *data mining* yang dilakukan sesuai dengan metodologi CRISP-DM yaitu dimulai dari *business understanding, data understanding, data preparation, modelling, evaluation* hingga *deployment.* Terdapat *deliverables* lain yang akan dihasilkan dari pelaksanaan proyek ini yaitu video presentasi, poster, dan *model deployment* yang disajikan melalui aplikasi heroku. *Deliverables* yang dihasilkan akan menyampaikan semua tahapan hingga hasil dari pengerjaan proyek ini. Berikut ini adalah rangkuman tahapan yang dilakukan dari awal hingga akhir pelaksanaan proyek.

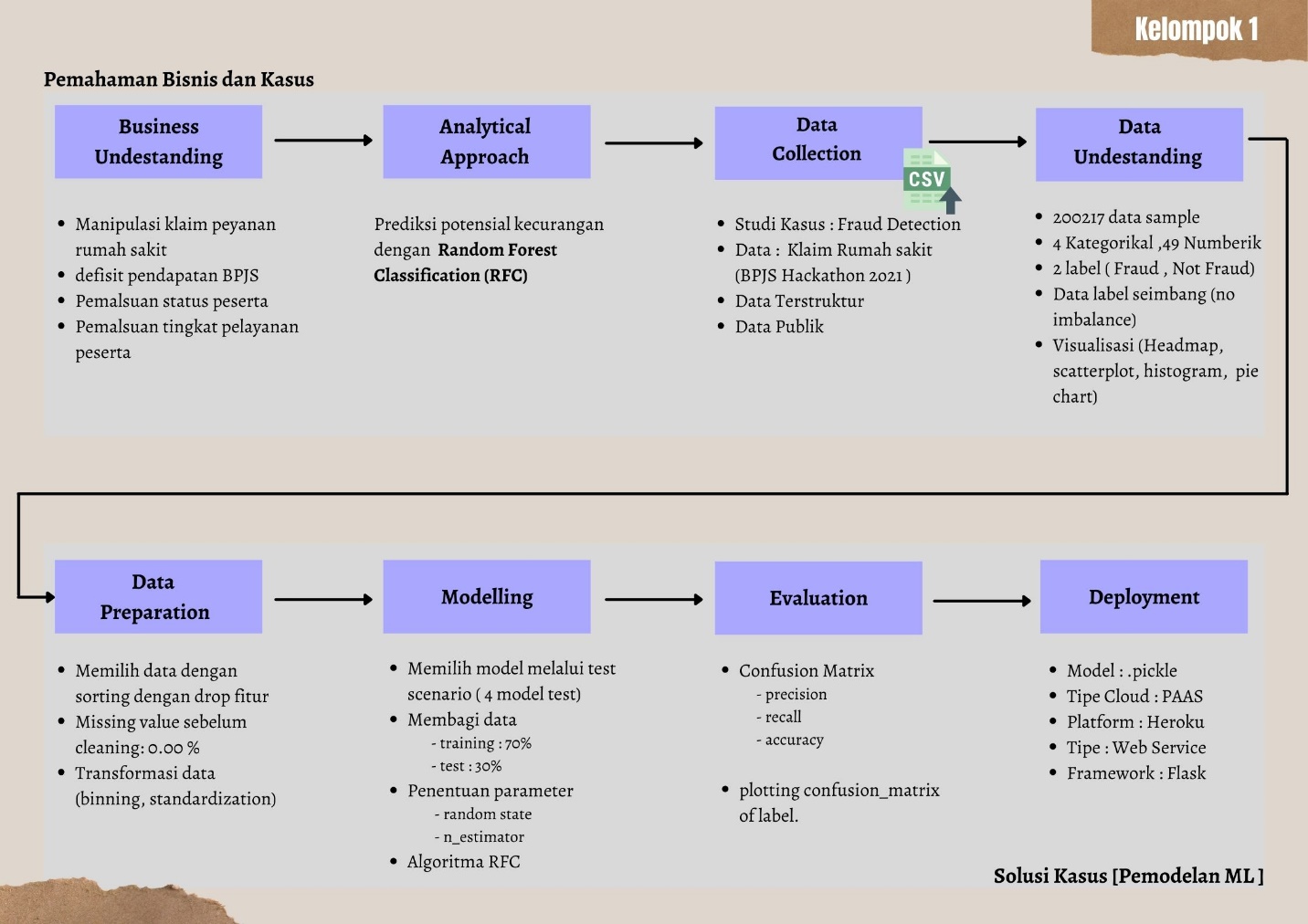
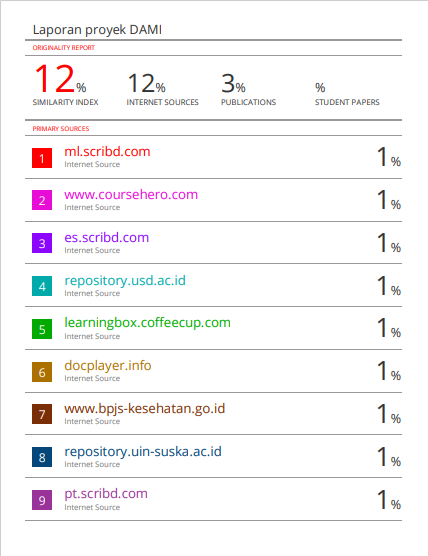


Figure 32 Rangkuman Tahapan Pengerjaan Proyek

# **LAMPIRAN**

Berikut merupakan tangkapan layar untuk hasil cek turnitin dari dokumen Laporan Akhir\_12S18018\_12S18019\_12S18064:

****